LAUR 96–145

Application of Variance-Based Methods to NUREG–1150 Uncertainty Analyses

Prepared by

Michael D. McKay
Statistics Group
Los Alamos National Laboratory

Prepared for

U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Washington, DC 20555

Under

NRC Job Code W6505, Task 5
Los Alamos National Laboratory
Harry F. Martz, Principal Investigator

June 28, 1996

Contents

List of Figures

Summary

This letter report is the product of Nuclear Regulatory Commission Job Code W6505, Task 5, a study of the feasibility and potential value of using variance-based methods as described in NUREG/CR–6311 (McKay, 1995) to supplement regression-based methods used in the probabilistic risk assessments of NUREG–1150. The regression-based methods of 1150 were used to assess prediction uncertainty and importance of inputs. This report shows that the Nuclear Regulatory Commission could assume a stronger position in support of future 1150-like analyses by augmenting the 1150 regression-based methods with more general variance-based methods that do not require an assumed form of a regression model. An examination of the 1150 methodology shows that the regression-based methods used therein are really a special case within the general framework of variance-based methods. Consideration of analysis objectives, the notion of importance, and concerns about the adequacy of the linear analysis model lead to the recommendation that general variance-based methods be used in conjunction with regression-based ones.

Introduction

In NUREG–1150 (1990), statistical analyses were directed towards quantifying uncertainty in model predictions associated with uncertainty in model input values. The different importance analyses with regard to model inputs constituted a large and complicated undertaking because of the sequence of computations necessary to move from initiating events through consequence calculations. One way to evaluate the potential for *variance-based methods*[1] to be a substantial improvement over the *regression-based methods* used in 1150 probabilistic risk assessments (PRAs) is to do a comparative analysis of results obtained from the two approaches. Because such an analysis would be very large and time consuming, this study undertakes conceptual and theoretical comparisons. Recognition of common objectives and quantitative definitions of importance can be used to make legitimate comparisons of analysis techniques. Failure to consider objectives and definitions has, in the past, lead to many arguments that certain measures or indicators of *sensitivity* and *importance* should be used or preferred over others. This report presents arguments and comparisons of 1150–like analysis methods as they arise from concisely stated definitions and assumptions.

In the 1150 PRAs, the assumptions under which statistical analyses were performed, for the most part, are the commonly used assumptions of a *linear analysis model*. Although analyses were often performed on rank-transformed data, the computations were those derived from linearity assumptions.[2] The methods themselves, which I call regression-based methods, form a cornerstone of statistical analysis. It is not being suggested that regression-based methods be discarded, for many of them have desirable features of robustness relative to their assumptions. However, departures from linearity can cause serious degradation of the power of regression-based methods. Therefore, validity of the assumption of a linear analysis model, whether with raw or rank-transformed data, is of paramount importance to evaluating how well objectives of the 1150 importance analyses are met.

Statistical analysis involves the application of laws of probability to observations for the purpose of finding plausible explanations of the observations consistent with the probability laws. When

[1] General variance-based methods also might be described as nonparametric regression methods. I am uncomfortable with the designation “variance based” because it suggests that regression methods are not variance based. Nevertheless, I continue to use the designation for convenience.

[2] The distinction between linearity in input variables and linearity in unknown parameters does not need to be made in this discussion.

using statistics, one makes assumptions about reality in order to define appropriate laws of probability from which hypotheses about the nature of observations may be formulated. This process involves construction of an *analysis model* which incorporates all of the assumptions about the source and nature of observations necessary to perform statistical analyses. The analysis model is formulated in such a way that the objectives of the analysis may be met. Sometimes there seems to be no practical alternative to the linear analysis model. When there is, however, this report strongly suggests that alternative techniques be included as part of a complete analysis methodology.

1. Determine a preliminary set of important input variables for expert elicitations.
2. Assess uncertainty importance of inputs using probability distributions obtained from expert panels.
3. Identify important initiating events.

Figure 1. Objectives of 1150 sensitivity and uncertainty analyses

Three objectives in the 1150 study are presented in Figure 1. First, in preliminary sensitivity studies, the importance of inputs relative to uncertainty in predicted (calculated) values is determined. Results from these preliminary studies are used to determine a set of input variables for expert elicitation. Second, the uncertainty importance of inputs is assessed using the distributions obtained from expert panels. Finally, the identification of important initiating events is made. This final objective in the 1150 study was not limited to the study of input variables. Of fundamental interest to the current study is the question “Would the additional use of variance-based methods put the Nuclear Regulatory Commission (NRC) in a stronger position in similar studies because of better quantification of uncertainty and importance of inputs?” This report answers the question in the affirmative based on theoretical considerations. However, details of implementation of such methods and computational demands remain unknown for 1150-like analyses. Therefore, this report considers strengths and weaknesses of both regression-based and variance-based methods as they might be applied in the PRAs of the 1150 study. Included in discussions that follow are considerations of details that would be necessary to carry out variance-based methods, including anticipated difficulties and possible approaches.

The remainder of the report is organized as follows.

- Definition of the 1150 PRA.
- Definitions related to uncertainty analysis and importance analysis.
- Some parts of an 1150 PRA where variance-based methods might be used to advantage.
- Mathematical description of regression-based methods.
- Mathematical description of general variance-based methods.
- Costs associated with regression-based and variance-based methods.
- Demonstration Application.
- Conclusions and Recommendations.

Short Description of the PRA Analyses of NUREG–1150

Descriptions of the PRAs and analysis methods performed in support of NUREG–1150 are presented in several places. Two particularly useful source are published in the open literature. An entire issue of *Nuclear Engineering and Design* in 1992 is devoted to the topic of thermal-hydraulics and related safety. The lead article by Breeding, Helton, Gorham, and Harper (1992a) provides a summary description of the PRA analysis methods. The next four articles discuss the four particular PRAs for the Surry Nuclear Power Station (Breeding, Helton, Murfin, Smith, Johnson, and Shiver, 1992b), the Peach Bottom Atomic Power Station, the Sequoyah Nuclear Plant and the Grand Gulf Nuclear Station, respectively. The second of the sources is a paper by Helton and Breeding (1993) in *Reliability Engineering and System Safety* which looks at the PRA methods in a somewhat more abstract and mathematical setting, thus providing many specifics of the computation procedures. The other documents upon which this report is based are the NRC reports NUREG–1150 and its supporting documents NUREG/CR–4550 (Ericson, Wheeler, Sype, Drouin, Cramond, Camp, Maloney, Harper, 1990), NUREG/CR–4551 (Gorham, Breeding, Helton, Brown, Murfin, Harper, and Hora, 1993), NUREG/CR–4551 (Breeding, Helton, Murfin, and Smith, 1990), and the LaSalle document, NUREG/CR–5305 (Brown, Payne, Jr., Miller, Johnson, Chanin, Shiver, Higgins and Sype, 1992). Material related to 1150 analyses appearing in this report has been extracted from all of these sources and from conversations with Frederick Harper and Ronald Iman at Sandia National Laboratories (SNL) Albuquerque, Jon Helton, Arizona State University (and an associate with SNL), and others. I will not make specific references to sources in the reference list except where it would be of particular value. My apologies in advance for any errors in attribution to the authors of the reports.

$$IE \rightarrowtail PDS \rightarrowtail APB \rightarrowtail STG \rightarrowtail cSTG \rightarrowtail rC$$

- IE denotes Initiating Events
- PDS denotes Plant Damage States
- APB denotes Accident Progression Bins
- STG denotes Source Term Groups
- cSTG denotes Consequences of Source Term Groups
- rC denotes Risk of Consequence

Figure 2. PRA calculation sequence

A matrix description of 1150 methods used by Helton and Breeding (1993) develops a sequence of conditional probability calculations. The probabilities refer to the sequence given in Figure 2. The probability calculations are represented in matrix notation as

$$rC = \mathbf{P}_x(IE)\mathbf{P}_x(PDS \mid IE)\mathbf{P}_x(APB \mid PDS)\mathbf{P}_x(STG \mid APB) \times cSTG$$

where, for example,

$$\mathbf{P}_x(IE) \text{ is a vector of elements Pr}(IE_i)$$

and

$$\mathbf{P}_x(PDS \mid IE) \text{ is a matrix of elements Pr}(PDS_i \mid IE_j) \ ,$$

whose dependence on a vector of input variables x is indicated. The elements of each conditional probability matrix are computed with codes such as SETS, TEMAC, EVNTRE, XSOR, MACCS and PRAMIS. At each step in a computation, codes use the outputs from preceding codes in the form of groups of events or states in *bins*. It is the binning or grouping of events and states that leads to the matrix description used. In addition, each code requires input values X for computation. The purpose of uncertainty analysis is to quantify and gain understanding of the variability in code output calculations due to variability in code input values x. The purpose of importance analysis is not only to identify, from among the inputs, the major contributors to uncertainty but also to estimate their contribution to uncertainty.

In the integrated analyses of NUREG–1150, the PRA in Figure 2 was evaluated for a sample of inputs on the order of 200 values selected by Latin hypercube sampling (LHS) (McKay, Conover, and Beckman, 1979). Thus, for each sample value X_i, the codes were run to produce a risk of consequence, denoted here as

$$rC_i = \text{ PRA}(X_i) \ .$$

In point of fact, the final consequence calculations are carried out for a sample of weather conditions and averaged to obtain rC_i. Finally, the 200 or so values are displayed in the form of a complementary cumulative distribution function (CCDF). The usual issues of uncertainty and importance, namely, those of assessing relative importance to inputs and subsets of inputs, are directed at the CCDFs.

The 1150 analyses become complicated because the questions asked of them—for example, “Which inputs are of significant importance to PDS probabilities?”—refer to a matrix of output values rather than to a single computed value or a simple time series output. The analyses are further complicated because of the subjective nature of binning of states, which has an unknown effect on final risk calculations. An important example of this effect comes up when considering how source term grouping affects consequence calculations in MACCS. In this example, the question has been asked by C Lui, of the NRC, and others as to whether STGs might be better treated in the “binning/sampling” way that weather conditions are treated. Although all of these issues are of importance, this report concentrates on basic questions related to fundamental analysis procedures used throughout most of the 1150 analyses.

There are two particular questions that seem very interesting but may be difficult to answer. First, we note that the probability matrices are really functions of the inputs X. That is, for example, one could write the conditional probability elements

$$\text{Pr}(PDS_i \mid IE_j) = p_{ij}(X) \ .$$

Moreover, even the choice of state and event partitioning with corresponding bins itself depends on values of the inputs X. That is, the i and j themselves could depend on X. Therefore, code outputs are likely to depend on some of the code inputs in a nonlinear fashion. Thus, appropriate measures of importance are not obvious. In particular, measures which depend on linearity assumptions are suspect. For example, the derivative of an output with respect to an input may not even exist because the binning effects make the relationship discontinuous.

The other question concerns the determination of the importance of initiating events. This question does not seem to be like ones concerning importance of inputs. A reasonable approach was taken in 1150 analyses and is described as follows. LHS samples of input values produced samples of risk values. The calculations were partitioned depending on initiating events (actually, the combination of initiating event and plant damage state). The mean value of risk was written as a sum of parts, either the "mean fractional contribution to risk (MFCR)" or the "fractional contribution to mean risk (FCMR)" corresponding to the states and events of interest (Breeding, Helton, Gorham, and Harper, 1992, p. 111). What seems to be missing in the literature is explanation and justification of the decompositions.

These two questions point to a primary issue of this report, namely, a suitable definition of "importance" for which importance measures and indicators may be derived and against which they may be judged. Because the regression-based methods of 1150 are special cases of variance-based methods (as will be shown later), objectives for importance analysis are discussed and evaluated in this report in a somewhat more general form relative to variance-based methods.

Uncertainty Analyses in NUREG–1150

The uncertainty analyses in 1150 are called analysis of *prediction uncertainty* by McKay (1995). A model prediction y is denoted as the result of the computation within a model $m(\cdot)$. That is,

$$y = m(x)$$

for a vector of input values x. Many predictions or model outputs were used in the course of the 1150 analyses. The ones I have been primarily associated with are consequence calculations from the computer code MACCS (Jow, Sprung, Rollstin, Ritchie, and Chanin, 1990). Therefore, much of what is said here may be overly influenced by the personal experience acquired from using MACCS.

The prediction uncertainty in y is determined by the triple

$$(f_x, V, m(\cdot))$$

where the inputs x take on values in V with probability (density) function f_x. The prediction uncertainty in y is characterized by its induced probability distribution f_y. In summary, prediction

$$\begin{gathered} x \sim f_x(x)\ ,\ x \in V \\ y = m(x) \\ y \sim f_y(y) \end{gathered}$$

Figure 3. Characterization of prediction uncertainty

uncertainty is determined and characterized in Figure 3. The objective of uncertainty analysis in 1150 is to estimate f_y from a set of runs, usually from an LHS sample of about 200.

Importance Analyses in NUREG–1150

Continuing to follow McKay (1995), the importance of inputs with regards to (prediction) uncertainty can be assessed by consideration of conditional probability distributions of the model output y conditioned on subsets S_x of the model inputs. Fundamentally, the importance of S_x depends on

$$\left\| f_y(y) - f_{y|s_x}(y \mid S_x = s_x) \right\| \tag{1}$$

which denotes a measure of the difference between the (marginal) distribution of y and the conditional distribution of y given the subset S_x. Because the marginal distribution of y can be written as

$$f_y(y) = \int f_{y|s_x}(y) f_{s_x}(s_x) ds_x\ , \tag{2}$$

it is argued that importance of the subset S_x is determined by the difference between f_y and the family of conditional distributions $\left\{ f_{y|s_x};\ s_x \in V_{s_x} \right\}$ indexed on the value of the subset S_x (McKay, 1995, p. 13–14). The notion that importance of an input subset is related to how well it controls the model prediction is reasonable. Intuitively, S_x is important if fixing its values substantially reduces the (conditional) prediction variance relative to the marginal prediction variance. With general *variance-based methods*, the prediction variance from the left hand side of Equation 2 is written in terms of the conditional variance from the right hand side, without any assumptions about the functional relation between y and s_x. Thus, it is reasonable that various (conditional) prediction variance ratios used in variance-based methods provide appropriate measures of importance.

In 1150, the differences in Equation 1 were investigated by examining the mean of the conditional distribution of y as a function of s_x. The functional dependence of y on a subset of input variables was determined by (stepwise) regression, often on the ranks of the variable values. The assumed form of the regression function is part of the issue of the analysis model which is discussed later in this report. Another method of comparing the conditional distributions with the marginal distribution of y was suggested by Iman and Hora (1990) in the analysis of TEMAC. They considered advantages of looking at differences in arbitrary quantiles of the distributions at the nominal values of the individual input subsets, which amounts to a local measure of importance. For general variance-based methods, importance is related to the expected value of the variance of the conditional distribution of y. This topic is examined in detail later in the report to show that regression-based methods are a special case of variance-based methods.

Places Where General Variance-Based Methods Can Be Used To Advantage

The following sections present four examples of areas where variance-based methods can be used to advantage. Complete arguments as to why variance-based methods provide an advantage follow from considerations developed later in this report in the sections defining regression-based and variance-based methods. Use of intuitive, undefined notions of importance are not sufficient for the arguments. The order of presentation of the ideas that follows is not meant to suggest priorities. It is just that many of the concepts addressed do not really align themselves in a hierarchical fashion.

Preliminary Screening The first place importance analysis was used in the 1150 PRAs was in *screening*. The purpose of screening is to decide which input variables can safely be *excluded* from further consideration as being potentially important. Before expert panels are presented lists of input variables for which probability distributions are needed, preliminary screening is used to arrive at a feasible number of inputs. Because substantial expense is involved with using expert panels for constructing probability distributions of inputs, an efficient screening process is very valuable. Obviously, the screening process must make some kinds of assumptions about the notion of importance for it to be effective. Thus, it is important that assumptions made for screening be consistent with the definitions and ultimate objectives of identifying inputs important to the uncertainty of consequence calculations. Regression-based methods can fail in preliminary screening by excluding an input that is really important to prediction variance because the output depends on the input in a nonlinear fashion. Variance-based methods can be expected to perform better, though it is not guaranteed to work without a sufficiently large sample size.

Base Line Studies with Generic Distributions With the screening exercises that precede expert panel work, there is a “chicken-or-egg” dilemma because importance of inputs depends on their probability distributions which are to be determined by the expert panel for important inputs. The reasonable approach taken in 1150 was to let code developers and other knowledgeable persons use the literature and their own experience to select preliminary distributions to use in the screening exercise. Problems with this approach are twofold. First, the group doing the screening actually defines the superset of inputs from which the screened set is chosen for the expert panels. Any bias introduced by the definition of a superset of inputs is a matter of concern but no more relevant to regression-based methods than to variance-based methods. However, the choice of sampling distribution used in screening may have more bearing on one method than another. In particular, breakdown of methods to identify important inputs—however importance might be defined—can have devastating effects in screening.[3]

An advantage that variance-based methods could have over regression-based ones is that variance can be estimated under different distributional assumptions using a single set of data using techniques developed by Beckman and McKay (1987). The techniques use re-weighting of observations depending on the input distribution to make inferences. On the other hand, if the linear analysis model assumptions hold, the constructed regression model should be independent of the probability distribution of the input variables. In this case, regression-based methods would be expected to work acceptably well.

[3] A study of possible sampling distributions to use for screening would be a good topic for research, possibly along the lines of generic maximum variance distributions.

Process Uncertainty and Process Evaluation Questions of importance and contribution to prediction uncertainty need not be limited to parameters in models, but may be directed to processes or phenomena both with and without specific reference to a model. Thus, questions of the importance of a submodel and the importance of weather phenomena, for example, are related. This point can be illustrated by considering the evaluation of the importance of the "weather process" to overall uncertainty of consequence calculations without modeling weather in the causative sense. By analogy, then, it can be possible to evaluate the importance of a submodel (calculation) through consideration of only its output calculations.

The effect of weather on consequence calculations has been described as *stochastic uncertainty* meaning that the *state* in which weather will be during a relevant time period around an accident is best described by a probability distribution. Therefore, a probability distribution can serve as a (descriptive) model for the weather process. Then, the questions that arise are how the weather distribution ought to be sampled and how its importance ought to be assessed. Both regression-based and variance-based methods could be used to evaluate *parameters* in a frequency distribution of weather types, if such a parameterized distribution is assumed. However, weather is usually described by a tabular empirical distribution for which an equivalent parametric form with a small numbers of parameters is not likely to exist. Therefore, it seems like the question of weather is better (and, possibly, exclusively) suited to variance-based methods, which do not rely on parametric representations and linear analysis models relating "weather" to consequence.

Process uncertainty can occur also at the interface between models where the output of one model is *binned* to become the input for the next model. An example of this is the STG step in Figure 2, which represents the interface between the models XSOR and MACCS in the calculation sequence. In the STG step, the output calculations of XSOR are binned or grouped according to assessed similarities in order to reduce the very large number of source terms that would need to be considered in MACCS. Effects of binning practices and assessment of binned output variables on uncertainty can be treated under process uncertainty. Thus, importance of and uncertainty contribution of binned code outputs used as inputs to the next code in a PRA and stochastic uncertainty are related issues. Because the output of a code can be viewed as the result of a general process, just as weather states are the result of the weather process, assessments with regards to source term groups, for example, would be made in the same way as assessments with regards to weather states. Just as in the case of weather, both regression-based and variance-based methods could be used to evaluate parameters in an assumed parametric form of a frequency distribution. However, source terms, like weather, are usually described by a tabular empirical distribution for which an equivalent parametric form with a small numbers of parameters is not likely to exist. Thus, general variance-based methods are preferable for assessment of importance.

Breakdowns It is known but not documented, as far as I can tell, that regression analysis has produced some unsatisfactory results when identifying important inputs. Any particular instances ought to be examined to see how variance-based methods might improve on regression-based methods.

Regression-Based Methods of NUREG–1150 and Their Analysis Model

The principal methods for assessing uncertainty importance used in the NUREG–1150 study are regression-based methods. This section begins the development that shows why regression-based methods are a special case of variance-based methods under the additional assumption of a linear analysis model. The presentation indicates those situations where regression-based methods are appropriate, although the appropriateness of regression-based methods is rarely tested in practice. The strength of regression-based methods comes from efficient estimation procedures because the assumption of a linear analysis model reduces the sample size required relative to that needed for general variance-based methods.

The models used in the NUREG–1150 study describe many phenomena where the relationship between input variables and the output is nonlinear. For the purpose of studying the uncertainty in model prediction, however, the 1150 models—TEMAC, EVNTRE, XSOR, and MACCS—were analyzed, by and large, by (linear) regression-based methods[4] of stepwise regression and partial correlation.[5] There is no question that these methods can be effective in identifying important input variables. However, the methods require for their validity the assumption of a linear relationship between model output and inputs to be at least approximately true. When that assumption breaks down, as it is known to do in some cases in the NUREG–1150 study, the process of identification of important input variables becomes suspect. Analysis calculations with rank data might lessen the impact of violation of the linearity assumption but they do not eliminate the difficulty. Unfortunately, the use of rank-transformed data also converts importance measures into much less quantitative importance indicators.

We now show why regression-based methods are a special case of variance-based methods with the additional assumption of a linear analysis model. Let the input (row) vector

$$x = (x_1, x_2, \cdots, x_p)$$

represent p inputs to a computer code used in the 1150 study. Let an output of the code be

$$y = y(t) \; .$$

The output y might be $y(t) = \text{Pr}(\text{Number of Early Fatalities} > t)$. If the computer code is represented by $m(\cdot)$, then we denote an actual code calculation or *computation model* by

$$y = m(x; t) \; .$$

The linear analysis model assumes that there is a (column) vector of unknown constants

$$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^t$$

such that the approximation to the computation model $m(\cdot)$,

$$\begin{aligned} x\beta &= \sum_{i=1}^{p} \beta_i x_i \\ &\simeq m(x) \; , \end{aligned}$$

[4] Although the phrase "regression-based methods" in this report refers to the commonly applied techniques of linear regression, statements extend in a natural way to nonlinear regression models.

[5] Analysis methods for the code TEMAC suggested by Iman and Hora (1990) for use with fault trees use R^2 from a polynomial regression as an "importance measure." The logarithm of the probability of the top event is the dependent variable.

The *linear analysis model* is

$$\begin{aligned} y &= E(y \mid x) + e \\ E(y \mid x) &= x\beta \end{aligned} \tag{3}$$

with

$$E(e) = 0$$

and, usually,

$$Cov[e, x] = 0\ .$$

Figure 4. Linear analysis model

is sufficient for the purposes of statistical analyses. It is to be understood that a constant term may be included in the model, for example, by introducing β_0 and $x_0 \equiv 1$.

The error term e in the usual linear analysis model in Figure 4 is treated as a random variable independent of x and having mean value zero. In the 1150 and similar studies, the error term is actually the difference between a code calculation $m(x)$ and the linear approximation $x\beta$. We return to this point later. For now, we use the usual linear analysis model with the random error term as an approximation to the computation model for the purpose of analysis. The phrase "*linearity assumption*" describes the analysis model in Equation 3.

In the linear analysis model, the variance of the output y is expressible as a linear combination of the variances and covariances of the inputs x. First of all, the variance of y in Equation 3 is given by

$$V[y] = V[x\beta] + V[e] + 2Cov[x\beta, e].$$

Under the (questionable) assumption that x and e are independent, the covariance term vanishes, leaving

$$V[y] = V[x\beta] + V[e]\ .$$

For independent components of x, this variance of y can be written as

$$V[y] = \sum_{i=1}^{p} \beta_i^2 V[x_i] + V[e]\ . \tag{4}$$

In general, though, not all of the inputs are independent. Therefore, the variance of y takes on a more complicated form which includes covariance terms,

$$V[y] = \sum_{i=1}^{p} \beta_i^2 V[x_i] + 2\sum_{i=1}^{p}\sum_{j<i}^{p} \beta_i\beta_j Cov[x_i, x_j] + V[e]\ . \tag{5}$$

In general vector/matrix notation, Equation 5 is written as

$$V[y] = \beta^t V[x]\beta + V[e]\ . \tag{6}$$

A reasonable measure of the importance of an input variable x_i which is independent of the other inputs is the term in Equation 4 corresponding to x_i. That term is $\beta_i^2 V[x_i]$ and, relative to $V[y]$, measures the contribution of input x_i to the variance of the output y using the linear analysis model. One estimates the variance component by way of a regression estimate of β, which we examine next. Before doing so, however, we mention two things. First of all, although there are other useful measures and indicators of importance, attention focuses here on the variance of the output. This perspective is sometimes referred to in the literature as *risk-reduction importance* or *uncertainty importance* (NUREG-1150, volume 1, page 3–10). Secondly, when the inputs are dependent, the construction and interpretation of importance measures and indicators is not as straightforward as for independent inputs but still possible. The partial correlation coefficient can be an example of an importance measure that can be used when inputs are not independent. We now continue with examination of regression-based methods.

In the 1150 study, the linear analysis model of Equation 3 was often used with only a subset of the input vector x of size s appearing in the $x\beta$ term. The effects of the remaining components of x were collected in the error term e. This is the case in stepwise regression, for example, where a subset of the inputs is selected to be important and to form the regression model.

Let S_x be a subset of inputs chosen in a stepwise regression. Let their subscripts be given by the set

$$I_s = \{i_1, i_2, \cdots, i_s\} \, .$$

Let the vector of the subset of the inputs be

$$x^s = (x_{i_1}, x_{i_{12}}, \cdots, x_{i_s}) \, ,$$

and let the corresponding vector[6] of βs be

$$\beta^s = (\beta_{i_1}, \beta_{i_2}, \cdots, \beta_{i_s})^t \, .$$

Then, the regression model that forms the basis for estimation and analysis is

$$y = x^s \beta^s + e^s$$

and it operates under the assumption that

$$\beta_i = 0 \text{ for } \; i \notin I_s \, .$$

Thus, the analysis implies that the subset regression model $y = x^s \beta^s$ is a reasonable approximation to computation model $m(\cdot)$. The β-vector is estimated via the usual least squares as $\widehat{\beta^s}$, and the variance of $x^s\beta^s$ can be estimated (with bias) with the estimator of β^s by

$$\widehat{V}[x^s\beta^s] = \widehat{\beta^s}^t V[x^s] \widehat{\beta^s} \, .$$

Also, from the regression, the variance of e^s can be estimated with the residual mean square. Therefore, the variance of y in Equation 6 from the stepwise regression model is estimated (with bias) by

[6] The notation for the "subset" β-vector properly might indicate a "subset" joint marginal distribution of x^s and y.

$$\widehat{V}[y] = \widehat{\beta^s}^t V[x^s]\widehat{\beta^s} + \widehat{V}[e^s]\,.$$

For the case of independent inputs, the estimated variance of y can be written as

$$\widehat{V}[y] = \sum_{k \in I_s} \widehat{\beta_k^s}^2 V[x_k] + \widehat{V}[e^s]\,.$$

In this form, it is seen that the term involving the square of the estimator of β might be replaced by a bias-corrected estimator. For this discussion, however, the biased estimator is sufficient.

All of this work finally gets us to a measure of importance from regression-based methods. The measure is an R^2 of the form

$$R^2 = \sum_{k \in I_s} \widehat{\beta_k^s}^2 V[x_k]/\widehat{V}[y]$$

which is similar to the usual R^2 used in regression, but takes into account the distribution of x.

In analyses with TEMAC within the 1150 study, Iman and Hora (1990), say that R^2 is a measure of uncertainty importance in the analysis of fault trees, but that it is not robust. They suggest using stepwise polynomial regression on the logarithm of y. Rather than using the ratio of variance estimates, they suggest using the ratio of quantile estimates because of the difficulty they have in obtaining stable variance estimates. Their paper does not present an argument as to why quantile estimates, even under the logarithm transformation, are stable. The authors may not have encountered any problems because they examined the ratios at a nominal value of each input rather than in expectation as is done with variance ratios. It is important to observe that their methods require an assumed polynomial model. This kind of assumption is not necessary for general variance-based methods.

In passing, we point out that if rank-transformed data are used in the stepwise regression, and the usual regression R^2 values are computed, an *indicator of importance* is created. We also point out that correlation and partial correlation coefficients are essentially variance ratios or R^2s. Thus, the discussion of the next section applies to these quantities, too.

The derivations of this section show that the process of finding a good subset regression via stepwise regression, for example, and using the R^2 from the regression as an importance measure for the subset is really variance based with the additional assumption of a linear analysis model. For the procedure to function properly, it is important that the xs be sampled appropriately, which was done with LHS in the 1150 analyses. We now examine regression-based methods to see where the methods are particularly strong and where they might break down.

Issues of Appropriateness of Regression-Based Methods

Assuming that a variance-based approach to importance satisfies analysis objectives, then a variance decomposition is a suitable measure of an input's importance. Thus, we have shown that regression-based methods are indeed suitable and proper methods as long as two assumptions are satisfied: (1) the linear analysis model is appropriate, and (2) the xs are properly sampled from their own probability distribution. The value of sampling xs for regression analysis of computer codes has been seen in practice. The preceding derivations help to explain why. We now address the linear analysis model assumption.

Without trying to be mathematically rigorous, the linear analysis model assumption combined with the added assumption that the error variance, $V[e]$, is constant and independent of x supports efficient estimation of the variance of the conditional expectation $x\beta$. Thus, one is really saving computer runs by making the assumptions. The problem, as expected, is that the linear analysis model is likely to be only approximately true, at best, and very far from true in many instances. To make matters worse, testing of the validity of the linear analysis model assumption seems to have been omitted in the 1150 analyses. The appropriateness of the linear analysis model is easily and properly questioned when considered against the computation models that were analyzed. In the NUREG–1150 study, the computation models $y = m(x)$ are highly nonlinear.

A result of a breakdown in the linear analysis model assumption is unreliable estimation of the variance decomposition and, thus, misleading indications or lack of indications of importances. Breakdowns are expected from three sources: (1) general nonlinearities in functional dependencies on the inputs x, (2) binning interfaces between codes, and (3) stochastic uncertainty. The question of the existence of other methods which do not rely on the linear analysis model is addressed in NUREG/CR–6311 (McKay, 1995). In the next section we see how these methods compare to regression-based methods and how they might be used to complement regression-based methods.

General Variance-Based Methods and Their Analysis Model

In the *general analysis model* in Figure 5, no particular assumptions are made about the form or distribution of the conditional expectation of y given S_x. For convenience, suppose $S_x = x$. Then, the basis of variance-based methods is the general analysis model and the well-known variance decomposition (Parzen, 1962),

$$V[y] = V[E(y \mid x)] + E(V[y \mid x]) \,. \tag{7}$$

The decomposition is seen to be the same one used in regression-based methods in Equation 6 without assumed linear form of the conditional expectation of y, required by regression-based methods, namely, $E(y \mid x) = x\beta$.

The objective of variance-based methods for 1150 analyses parallels that of regression-based methods, and is to find input variable subsets S_x such that

$$E(y \mid S_x) \simeq m(x)$$

is a reasonable approximation to the computation model $m(\cdot)$.

For any arbitrary subset S_x of inputs, the *general analysis model* is

$$y = E(y \mid S_x) + e$$

with

$$E(e) = 0$$

and

$$Cov[E(y \mid S_x), e] = \Sigma \ .$$

Figure 5. General analysis model

In integral representation, with x standing for a variable subset, Equation 7 becomes

$$\begin{gathered} \int (y-\mu_y)^2 f_y(y)dy \ = \\ \int (E(y \mid x)-\mu_y)^2 f_x(x)dx \ + \ \int\int (y-E(y \mid x))^2 f_{y|x}(y) f_x(x) dy dx \ . \end{gathered} \tag{8}$$

This latter representation may make it clearer how the variance decomposition works. The variability in the random variable y is

$$V[y] = \int (y-\mu_y)^2 f_y(y)dy \ .$$

The amount of variability "explained" by another random variable x is measured by the size of the variance of the expected value of y conditioned on x,

$$V[E(y \mid x)] = \int (E(y \mid x)-\mu_y)^2 f_x(x)dx \ .$$

McKay (1995) points out how estimation of Equation 7 parallels classical analysis-of-variance (AOV). The integral form of Equation 8 suggests that for estimation via the linear analysis model, the variance decomposition parallels the AOV for regression analysis,

$$\begin{gathered} \text{Total sum of squares } (\text{SST}) = \\ \text{Regression sum of squares } (\text{SSR}) + \text{ Error sum of squares } (\text{SSE}) \ . \end{gathered}$$

The parallel is seen by the substitutions

$$\begin{gathered} \int \rightarrowtail \sum \\ \mu_y \rightarrowtail \overline{y} \\ E(y \mid x) \rightarrowtail x\widehat{\beta} \ . \end{gathered}$$

When the variables x are sampled appropriately by LHS, for example, the three sums of squares, SST, SSR and SSE, can be used to estimate the terms in Equation 7.

Finally, the measure of importance of x with variance-based methods is the *correlation ratio*,

$$\eta^2 = \frac{V[E(y \mid x)]}{V[y]} \ .$$

Regression-Based Methods Are a Special Case of Variance-Based Methods

That regression-based methods are a special case of variance-based methods means that the R^2 from a regression model used to assess importance is really an estimator of the correlation ratio given the regression model. That is,

$$\begin{aligned} R^2 &\simeq \frac{\beta^t V[x]\beta}{V[y]} \\ &= \eta^2 \end{aligned}$$

for the regression model $E(y \mid x) = x\beta$.

When looking at a single input variable x, the correlation ratio from a linear regression model is equivalent to the square of the *correlation coefficient* ρ, defined by

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \,.$$

Because

$$\begin{aligned} \sigma_{xy} &= \text{Cov}(x, y) \\ &= \text{Cov}(x, E(y \mid x)) \\ &= \text{Cov}(x, x\beta) \\ &= \beta\sigma_x^2 \,, \end{aligned}$$

we see that,

$$\rho^2 = \eta^2 = \beta^2 \frac{\sigma_x^2}{\sigma_y^2} \,.$$

Therefore, for a single input variable and in stepwise linear regression, the R^2 of a model is an estimate of the correlation ratio of variance-based methods under a linear analysis model.

For the linear analysis model in the univariate case, we see that

$$\begin{aligned} \beta &= \frac{\sigma_{xy}}{\sigma_x^2} \\ &= \rho \frac{\sigma_y}{\sigma_x} \,, \end{aligned}$$

which extends to the multivariate case as

$$\begin{aligned} \beta &= V[x]^{-1}\sigma_{xy} \\ \beta^t V[x]\beta &= \sigma_{xy}^t V[x]^{-1}\sigma_{xy} \\ &= \eta^2 \times \sigma_y^2 \,. \end{aligned}$$

Costs and Strengths of Regression-Based and Variance-Based Methods

One may not have an option when it comes to choice of methods for importance analysis. Nevertheless, it is important to know the possible costs associated with particular procedures. The next two sections summarize main points.

Regression-Based Methods Regression-based methods require many fewer computer runs than does variance-based methods. However, the relationship that

$$\rho^2 \leq \eta^2 \ ,$$

found in Kendall and Stuart (1979), shows that regression-based methods can miss important inputs that variance-based methods can find because ρ^2 and R^2 can be small, indicating lack of importance, while η^2 can be large, properly indicating importance. Thus, it is shown that regression-based methods can fail in preliminary screening and elsewhere to find important inputs.

Variance-Based Methods Under the linear analysis model in Figure 4, the form of the conditional expectation of y makes its estimation apparent via estimation of beta. For the general analysis model in Figure 5, the conditional expectation of y depends on X in an unspecified manner. Therefore, the conditional expectation and its variance are estimated via sampling theory. The number of computer runs necessary to assure adequate estimation is unknown in the general case because it will depend on the specific model $m(\cdot)$ under study. It seems reasonable, however, to assume that the number required is proportional to that required under a linear analysis model, with the constant of proportionality being related to the complexity of $m(\cdot)$. The complexity of $m(\cdot)$ might be indicated by the number of parameters in a suitable[7] Taylor series expansion of $m(\cdot)$. Therefore, variance-based methods might need many computer runs to properly identify important inputs.

Summary The main point of this discussion is not whether R^2 works as a measure of importance, but that the source of breakdown of regression-based methods occurs because of breakdown of the linear analysis model assumption. Thus, the strength of variance-based methods is not so much in the use of variance ratios and R^2 as it is in the use of a general analysis model. When the linear analysis model is valid, R^2 can be estimated efficiently from a linear model. When the linear analysis model is not valid, general variance based-methods must be used. Therefore, this report recommends that variance-based methods be used in addition to regression-based methods, to provide the NRC decision makers with a better foundation and better quality analytical support for making decisions that depend on proper assessment of the importance of input variables in PRA analysis codes.

Demonstration Application

In the following demonstration application, several points relevant to measuring importance are discussed in relation to a model with two input variables. The model output is a continuous function of one of the inputs but a discontinuous one of the other. The response to the second input is intended to represent what might be encountered with a model which uses (discontinuous) weather regimes or categories as an input. Results from a simulation study are presented as examples of what one might encounter in practice.

[7] The notion of suitability and precise definitions of "suitable" need to be developed before these ideas can be used.

Importance Indicators As shown earlier, regression-based methods are a subset of general variance-based methods. In variance-based methods, the importance of an input x is measured by the correlation ratio,

$$\eta^2 = \frac{V[E(y \mid x)]}{V[y]} .$$

Regression-based methods assume that the functional form of the conditional expectation of y, which appears in the numerator of the expression for η^2, is known. In the linear case, the expectation is written often as

$$E(y \mid x) = x\beta .$$

When a linear function for the conditional expectation of y is assumed, the correlation ratio is equal to the square of the (multiple) correlation coefficient,

$$\begin{aligned} \eta^2 &= \frac{\beta^t V[x]\beta}{V[y]} \\ &= \rho^2 . \end{aligned}$$

Estimation of Importance Indicators The correlation ratio and the correlation coefficient are estimated (with bias) from a sample of values as follows. Let

$$\{x_i,\ i = 1, \cdots, n\}$$

be a random sample of size n from f_x, and let

$$\{y_{ik},\ k = 1, \cdots, r\}$$

be a conditionally independent random sample of size r from $f_{y|x_i}$ for $i = 1, \cdots, n$. Let the sample means be

$$\overline{y}_i = \sum_{k=1}^{r} y_{ik} , \quad \overline{y} = \sum_{i=1}^{n} \overline{y}_i , \quad \text{and} \quad \overline{x} = \sum_{i=1}^{n} x_i .$$

An analysis of variance decomposition of sums of squares is given in Figure 6. Estimates of the correlation ratio and the square of the correlation coefficient are indicated at the bottom of the figure. An advantage of using the estimators indicated is that they have the property of their population counterparts that $0 \leq \widehat{\rho}^2 \leq \widehat{\eta}^2 \leq 1$, as implied in the figure. A disadvantage is that bias is introduced because of the covariance structure of the y_{jk} induced by the sample design. For example, the expected value of the total sum of squares is, approximately,

$$\begin{aligned} E(\mathrm{SST}) &= r(n-1)V[y] + (r-1)E(V[y \mid x]) \\ &= r(n-1)\sigma_y^2 + (r-1)\overline{\sigma}_e^2 , \end{aligned}$$

which can be driven to $nr\sigma_y^2$ with n for fixed r. Similarly,

$$\begin{aligned} E(\text{SSB}) &= r(n-1)V[E(y \mid x)] + (n-1)E(V[y \mid x]) \\ &= r(n-1)V[E(y \mid x)] + (n-1)\overline{\sigma}_e^2\,, \end{aligned}$$

and

$$\begin{aligned} E(\text{SSW}) &= n(r-1)E(V[y \mid x]) \\ &= n(r-1)\overline{\sigma}_e^2\,. \end{aligned}$$

With this caution in mind, we proceed to describe the demonstration model.

Source of Variation	Degrees of Freedom	Sum of Squares
Total	$nr-1$	$\text{SST} = \sum_{i=1}^{n}\sum_{k=1}^{r}(y_{ik}-\overline{y})^2$
Between	$n-1$	$\text{SSB} = r\sum_{i=1}^{n}(\overline{y}_i-\overline{y})^2$
Regression	1	$\text{SSR} = r\left[\sum_{i=1}^{n}(\overline{y}_i-\overline{y})(x_i-\overline{x})\right]^2 / \sum_{i=1}^{n}(x_i-\overline{x})^2$
Error (lack of fit)	$n-2$	SSE = SSB − SSR
Within	$n(r-1)$	$\text{SSW} = \sum_{i=1}^{n}\sum_{k=1}^{r}(y_{ik}-\overline{y}_i)^2$

$$\begin{aligned} \widehat{\eta}^2 &= \text{SSB/SST} \\ \widehat{\rho}^2 &= \text{SSR/SST} \end{aligned}$$

Figure 6. Analysis of variance decomposition of sums of squares

The Model The demonstration application is designed to show the range of results of analyses that can be encountered for continuous inputs and categorical inputs and for regression-based and general variance-based methods.[8] The model used is simply a randomly selected polynomial function,

$$\begin{aligned} y &= m(x,d) \\ &= \text{Legendre polynomial in } x \text{ of degree } d\,, \end{aligned}$$

with

$$\begin{aligned} x &\sim \text{ Uniform on } [-1,+1] \\ d &\sim \text{ Uniform on } \{1,2,3,4,5\}\,. \end{aligned}$$

Thus, with probably $1/5$, y is a Legendre polynomial in x of degree d. The different polynomials represent 5 different responses of y to x. An example of 5 categorical responses might be 5 different weather categories. The responses of y as a function of x for the 5 polynomials are plotted in Figure 7. Legendre polynomials have the properties that they are orthogonal and integrate to 0 on the interval $[-1,1]$. The mean value of y is 0, and its variance is $3043/17325 \simeq 0.1756$.

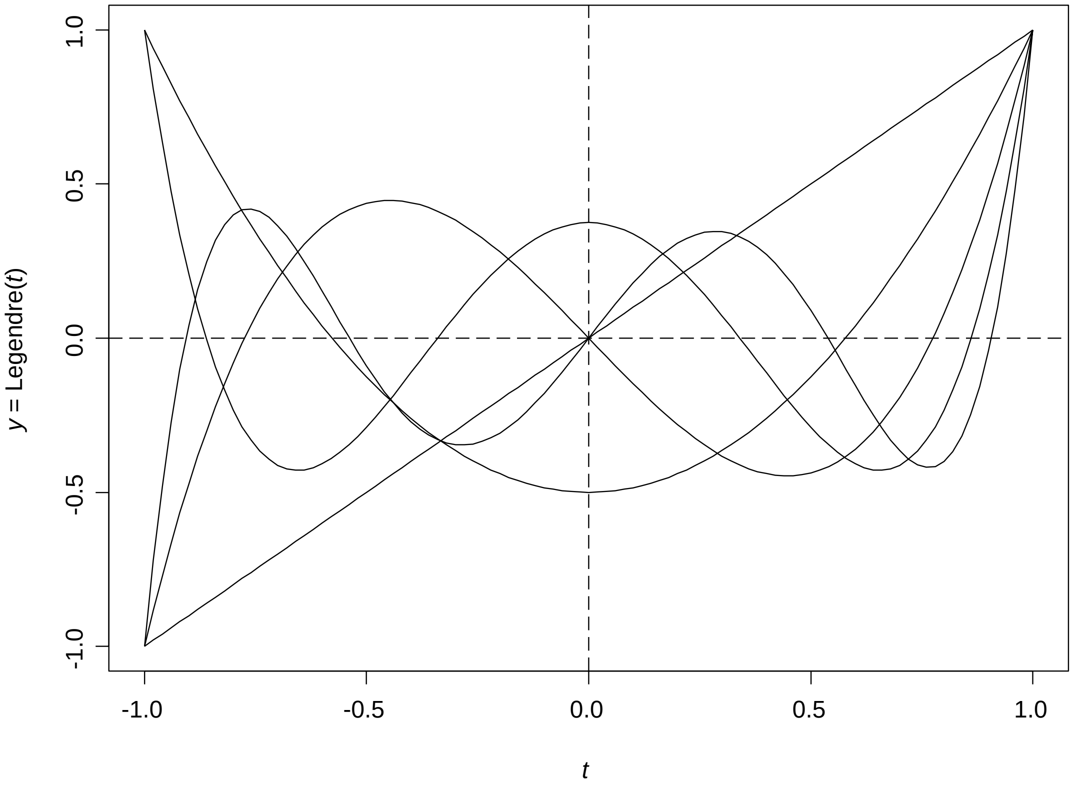

Figure 7. Legendre polynomials of degree 1 to 5

[8] Consideration of alternative measures of association for categorical variables are not included in this report.

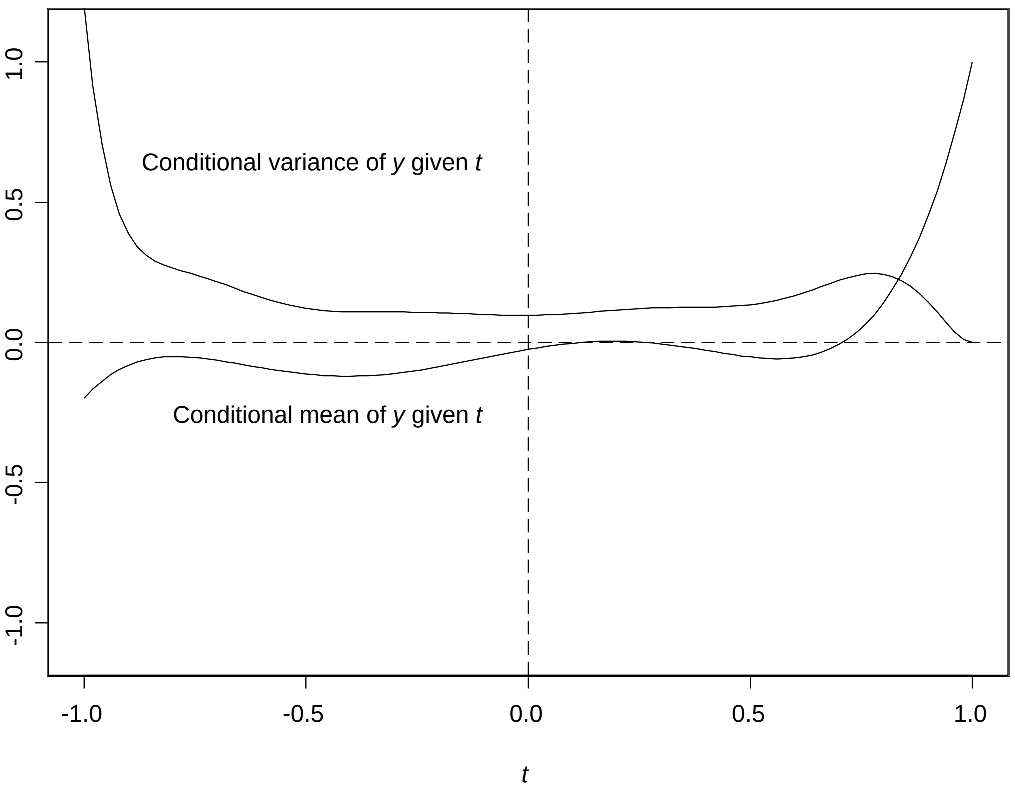

Figure 8. Conditional mean and variance of y given x

Importance of d and x Figure 7 illustrates the "importance" of x as the change in y while moving along one of the 5 curves. The "importance" of d is indicated as the change in y as one moves among the 5 curves in a vertical direction for fixed x. The meaning of importance suggested is the change in conditional mean of y, which in now presented.

For $d \in \{1, 2, 3, 4, 5\}$, the mean and variance of y conditioned on d are given by

$$\begin{aligned} E(y \mid d) &= 0 \\ V[y \mid d] &= \frac{1}{2d+1} \,. \end{aligned} \tag{9}$$

The conditional mean and variance of y given x are shown in Figure 8. The mean is a smoothly varying, generally increasing function of x. The variance is maximum at $x = -1$ and is 0 at $x = +1$. The variance is roughly constant between –0.5 and 0.5.

Importance indicators, whether they are η^2 or ρ^2, are intended to show how an input influences, in some way, the value of an output. In variance-based methods, we look to see how an input controls the value of the output in the sense of how much the variance of the output is reduced when the value of the input is held fixed. Specifically, η^2 measures how much, on average, the variance of y conditioned on an input is reduced relative to the unconditional variance. Equivalently, η^2 measures how closely the variance of the conditional expected value of y matches, on average, the unconditional variance of y.

As an example, we see from Equation 9 that, for all values of d, the expected value of y given d is a constant, namely, 0. Thus, the variance of the conditional expected value of y is 0 and not close

TABLE I
Importance indicators with $y =$ Legendre$(x; d)$ for
d uniform on $1, 2, \cdots, d_{max}$ and x uniform on $[-1, 1]$.

	d		x	
d_{max}	ρ^2	η^2	ρ^2	η^2
Values for model used in demonstration application				
5	0	0	.08	.20
Values for a range of polynomial models for comparison				
1	—	—	1.0	1.0
2	0	0	.31	.50
3	0	0	.16	.33
4	0	0	.11	.25
5	0	0	.08	.20
10	0	0	.03	.10

to the unconditional variance of y. Therefore, d is measured to be unimportant by variance-based methods, in general, including regression-based methods.

Population values of η^2 and ρ^2 are given in Table I, and components of the correlation ratio are found in Table II. We see from the tables that the expected value of y given d does not change. Therefore, both the correlation ratio and the correlation coefficient are 0, which indicates not only that d is unimportant, but that it is an irrelevant input with respect to the importance measure. However, while it is true that the expected value of y given d is constant for all d, the variance of y given d is not, as shown in Equation 9. Therefore, one is lead to the observation that variance-based measures of importance, which use the variance of the conditional expectation of y, may not always be effective as indicators of importance.[9]

Sampling Variability of Importance Measures Sampling variability of estimators of the correlation ratio and the correlation coefficient for x and d can be significant, as shown in this section. Estimators of the importance indicators were computed in a simulation study of size 100 for several values of n and r.

For inputs with an infinite number of values, like the continuous input x, n is the number of distinct values sampled for each simulation run and r is the number of replicate values of y obtained by sampling the other inputs. For inputs like d, which can take on only a finite number of values, the number of distinct values can be less than n. In this demonstration, the number of distinct values of d is always 5. Therefore, for d, the number of replicate values of y is $r \times n/5$, which was made an integer by the choices of r and n.

[9] Conditional entropy might be evaluated as an alternative measure of importance in this case.

TABLE II
Variance decomposition

Input	$E(y \mid \text{input})$	$V[E(y \mid \text{input})]$	$E(V[y \mid \text{input}])$	$V[y]$
d	0	0	$\frac{3043}{17325} \simeq 0.1756$	$\frac{3043}{17325} \simeq 0.1756$
x	polynmial in x (See Figure 8)	$\frac{3043}{86625} \simeq .0351$	$\frac{12172}{86625} \simeq 0.1405$	$\frac{3043}{17325} \simeq 0.1756$

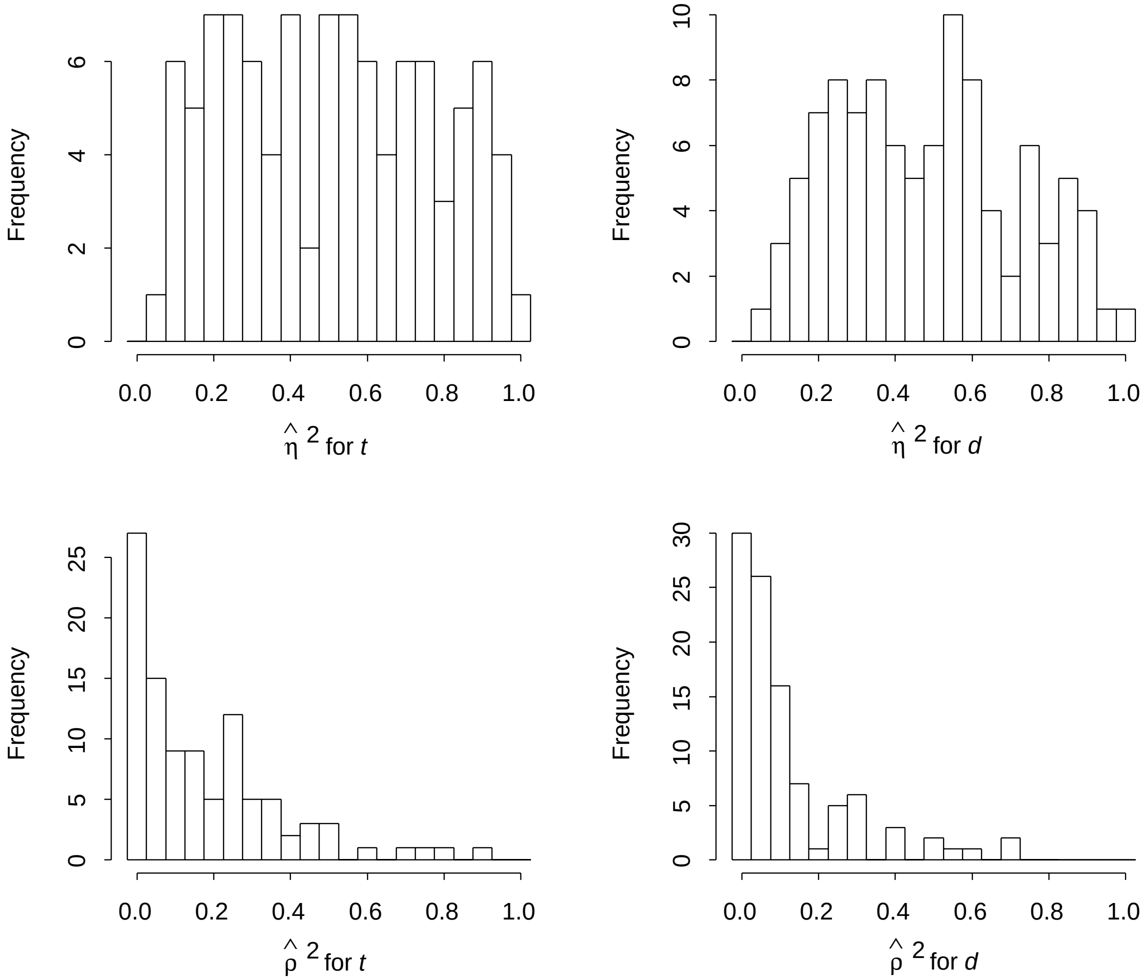

Figure 9. 100 simulations with $n = 5$ and $r = 2$

The minimum number of replicated values required for estimation is $r = 2$. Figures 9, 10, and 11 show histograms of the estimators of the correlation ratio and the square of the correlation coefficient

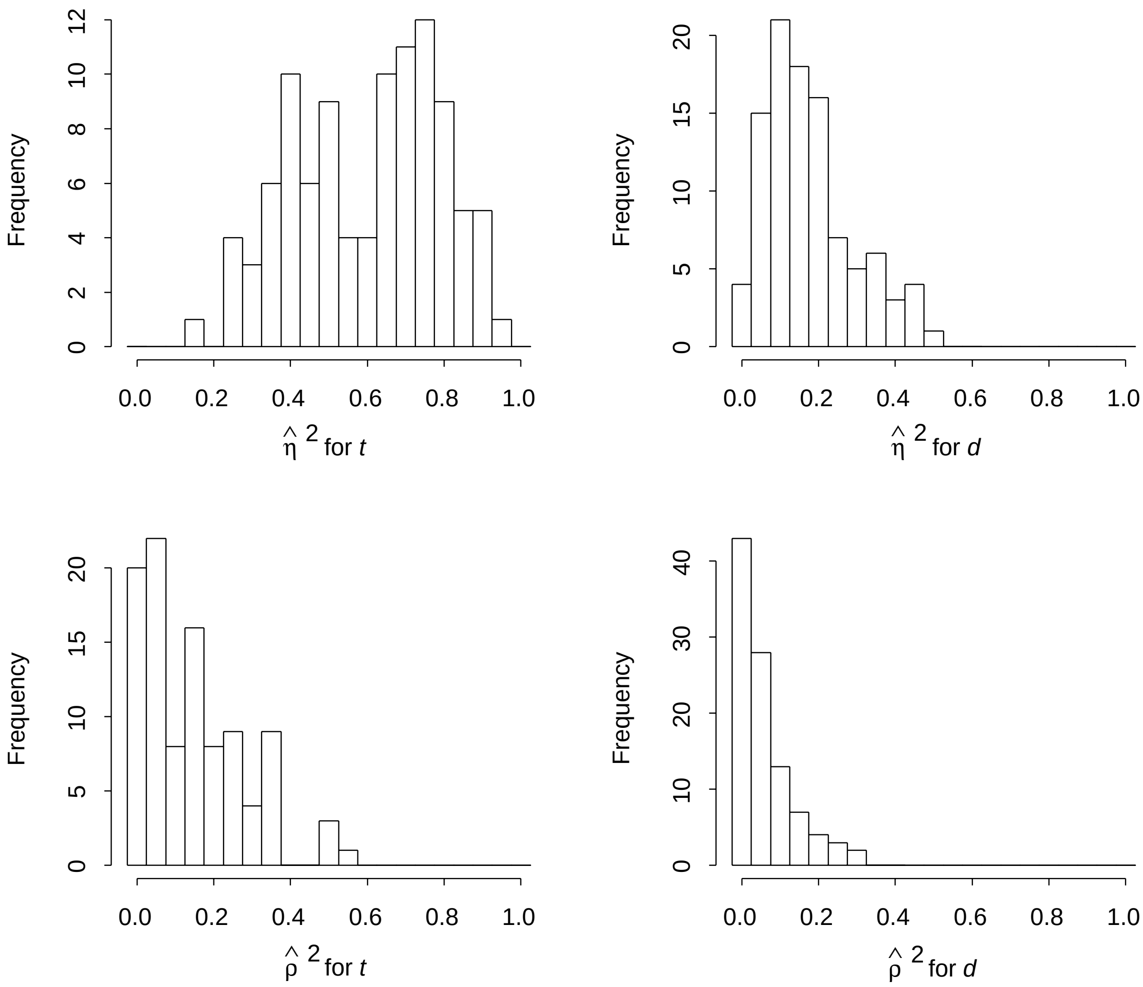

Figure 10. 100 simulations with $n = 10$ and $r = 2$

(both defined in Figure 6) for $r = 2$ and $n = 5, 10, 50$.

The upper left graph in each of the figures is a histogram of the estimator of the correlation ratio for x. The estimates of η_x^2 have vary large spreads for $n = 5$ and $n = 10$ which converge to a biased estimate as indicated by their distribution for $n = 50$. To get an idea of the bias, one can examine the expected values of the sums of squares in the estimator of η^2. The ratio of expectations is given by

$$\frac{E(\text{SSB})}{E(\text{SST})} = \frac{(n-1)(r-1)\eta^2 + (n-1)}{rn - 1 - (r-1)\eta^2} .$$

In the limit with n,

$$\lim_{n\to\infty} \frac{E(\text{SSB})}{E(\text{SST})} = \left(1 - \frac{1}{r}\right)\eta^2 + \frac{1}{r}$$

While the ratio of expectations is not equal to the expectation of the ratio, the result shows the order of the bias with r in the estimator of η^2. In this demonstration, for which $\eta_x^2 = 0.2$, the limiting value of the ratio of expectations is 0.6, clearly consistent with Figure 11. Figure 12 indicates more succinctly for $r = 2$ the convergence of the biased estimator with n. To complete the picture, Figure 13 shows the convergence with r of the estimator to the population value of 0.2 for $n = 50$.

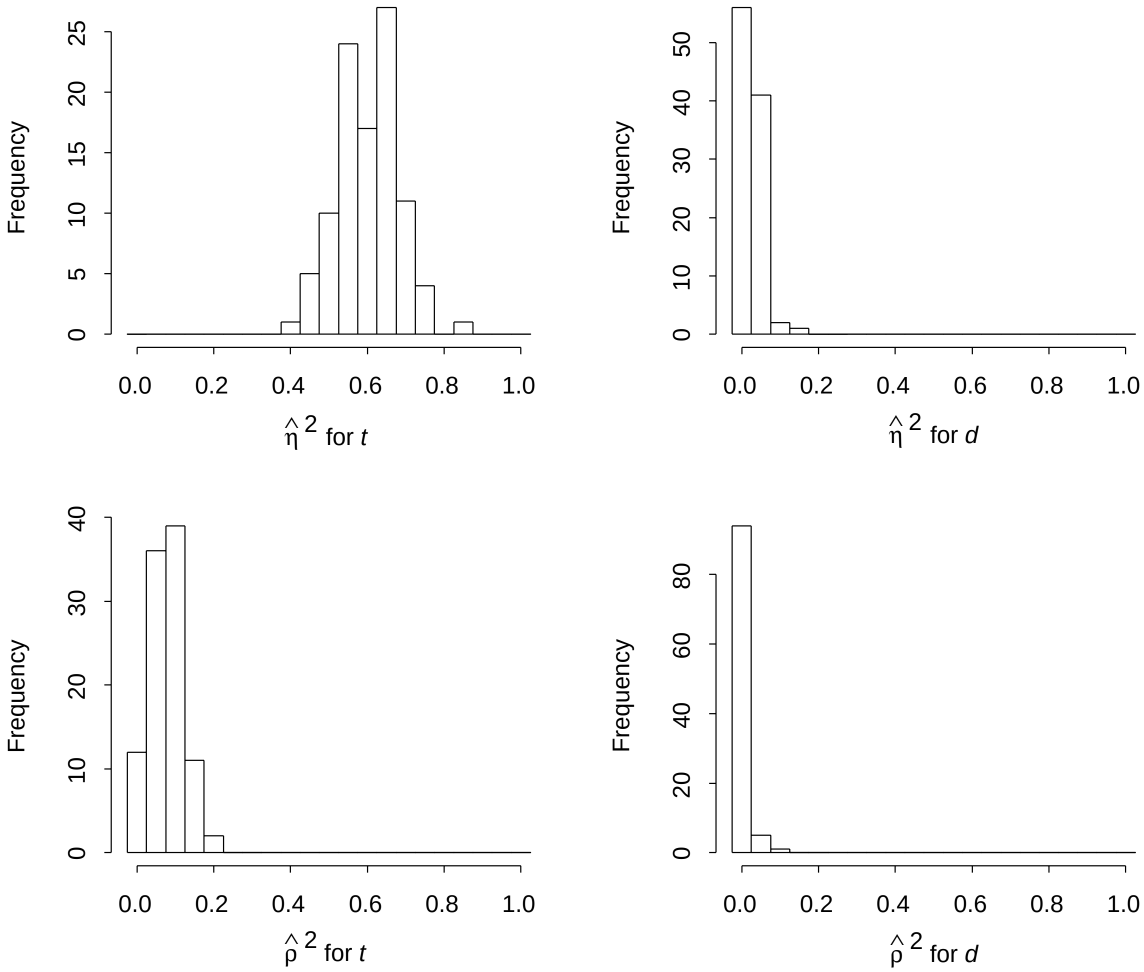

Figure 11. 100 simulations with $n = 50$ and $r = 2$

The upper right graph in each of the Figures 9, 10, and 11 is a histogram of the estimator of the correlation ratio for d. The population value of the correlation ratio for d is 0. Therefore, we compare the simulation results with the limit with n with $\eta^2 = 0$ of

$$\lim_{n\to\infty} \frac{E(\text{SSB})}{E(\text{SST})} = \lim_{n\to\infty} \frac{n-1}{rn-1} = \frac{1}{r} \,,$$

which is not consistent with the results of Figure 11. These results indicate that the ratio-of-expectations approximations is not valid for very small values of the correlation ratio. This fact poses no problems in application.

The lower left and right graphs in the figures pertain to estimators of the square of the correlation coefficient. Results for $n = 5$ indicate, as in the case of the correlation ratio, that the sample size is to small for meaningful conclusions. Figure 11 results are consistent with theoretical results for the correlation coefficient, that say that neither x nor d is an important input.

Finally, estimates of η_x^2 and ρ_x^2 for $n = 5, 10, 50$ grouped by $r = 2, 5, 20$ are shown in Figures 14 and 15, respectively. The figures support the conclusions that a substantial number of computer runs may be necessary to adequately estimate the correlation ratio, and that the correlation coefficient can fail to detect important inputs.

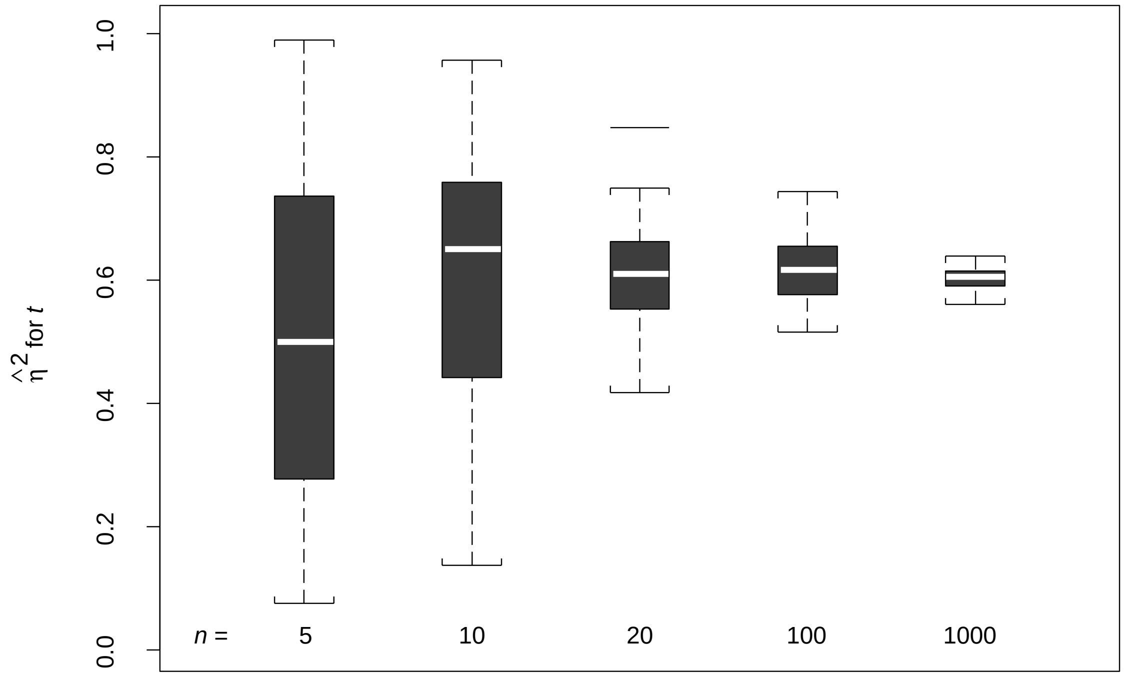

Figure 12. Convergence of the estimated correlation ratio for x with $n = 5, 10, 50, 100, 1000$ to biased value for $r = 2$, where each box plot contains 100 simulations

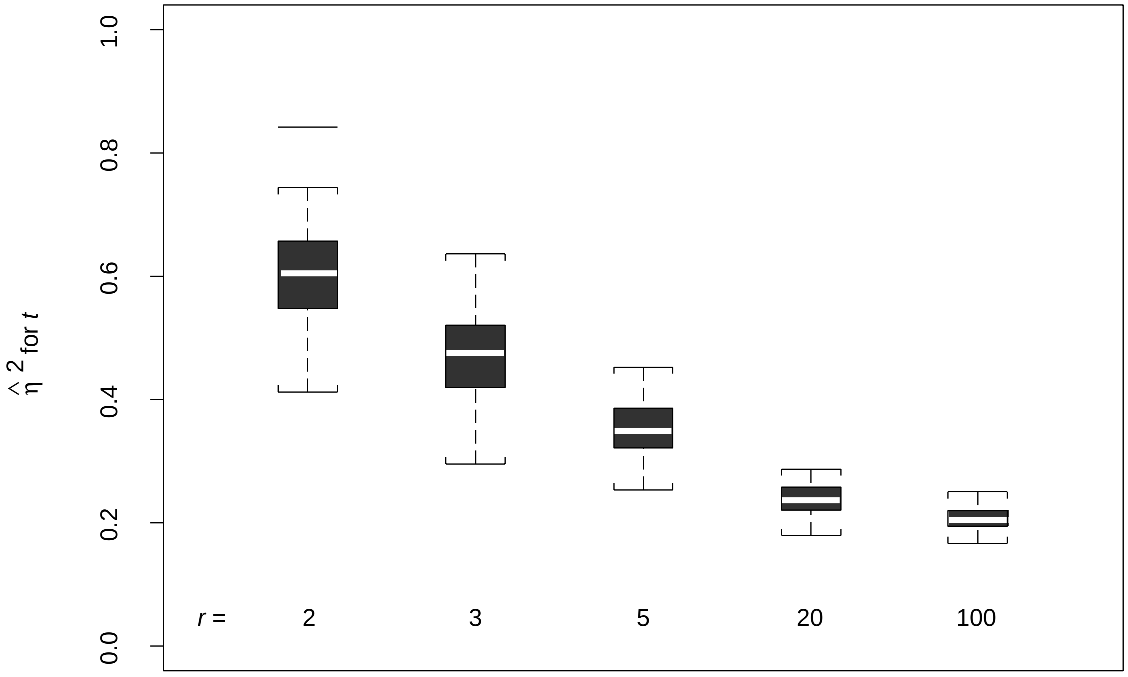

Figure 13. Convergence of the estimated correlation ratio for x with $r = 2, 3, 5, 20, 100$ for $n = 50$, where each box plot contains 100 simulations

Conclusions and Recommendations

The NRC could assume a stronger position in support of future 1150-like PRAs by augmenting regression-based methods of importance analysis with more general variance-based ones. General

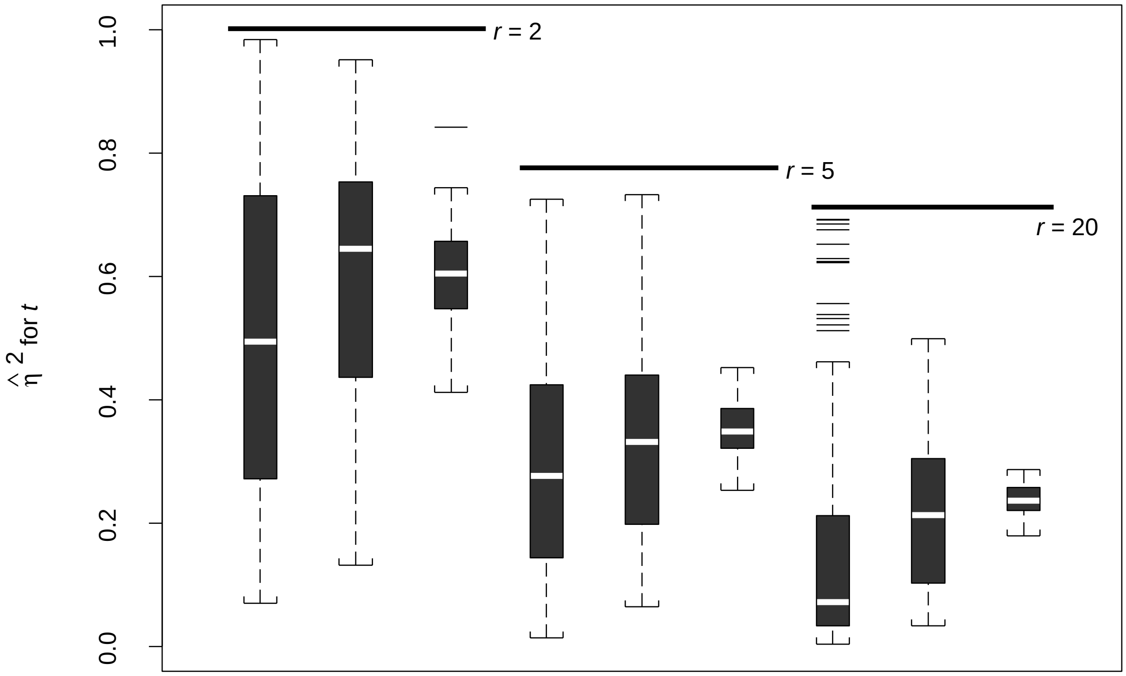

Figure 14. Estimates of the correlation ratio for x with $n = 5, 10, 50$ grouped by $r = 2, 5, 20$

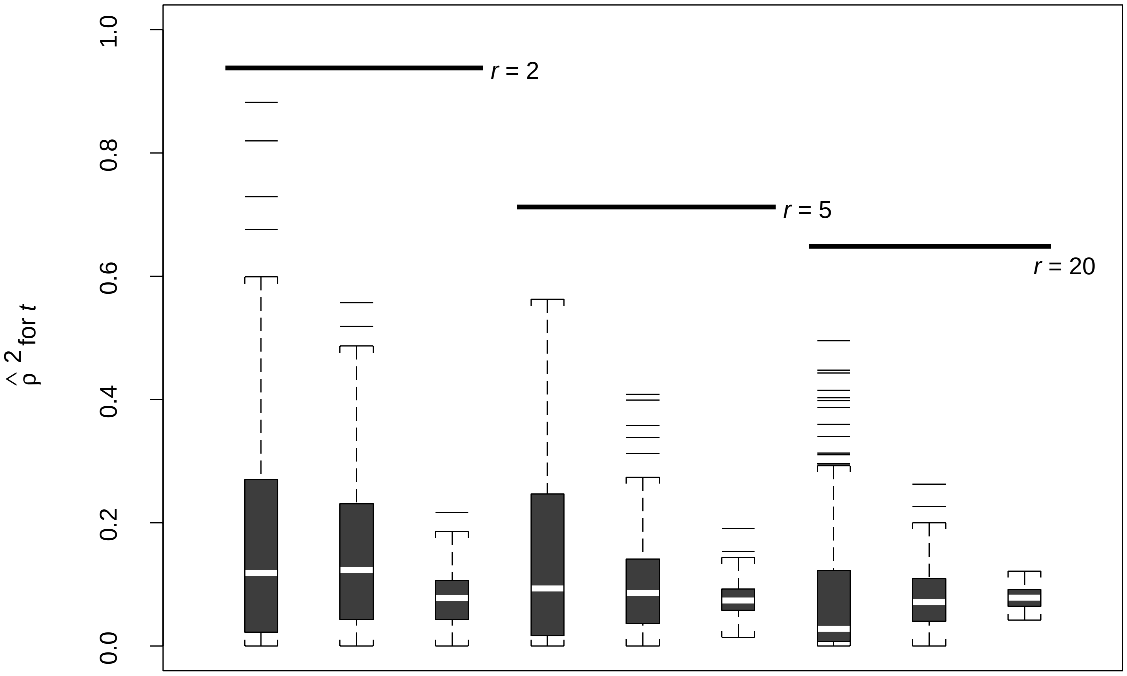

Figure 15. Estimates of the square of the correlation coefficient for x with $n = 5, 10, 50$ grouped by $r = 2, 5, 20$

variance-based methods are theoretically superior to regression-based methods which can be expected to fail to detect important inputs in some situations when models are nonlinear. However, general variance-based methods require substantially more computer runs to be effective. Nevertheless, through the use of general variance-based methods together with regression-based ones, NRC decision makers would have sound theoretical foundation and better quality analytical support for making decisions that depend on proper assessment of the importance of input variables in PRA analysis codes.

Several recommendations and topics for investigation and research are indicated in this report.

- Whether regression-based methods or variance-based methods are used, techniques for evaluating the adequacy of assumptions of the analysis model, particularly, in the case of regression-based methods, and adequacy of the sampling design, particularly, in the case of variance-based methods, should be incorporated in analysis procedures.
- Methods for evaluating adequacy, possibly along the lines of goodness of fit procedures and cross validation, should be researched and developed for incorporation into PRA uncertainty and importance analysis procedures.
- Efficient sampling plans for use with general variance-based methods should be developed in order to reduce the computational requirements of variance-based methods. A possible starting point would be the investigation of reuse or multiple-use sampling plans for LHS.
- Because of the cost effectiveness of screening exercises for assessing importance of input variables, study of appropriate sampling distributions to use with screening would be a good topic for research, which might begin along the lines of using generic maximum variance distributions.
- Many important issues related to stochastic uncertainty and the use of binning in PRAs remain undiscovered or unresolved. Research into either or both of these areas is recommended.

References

Beckman, R. J. and McKay, M. D. (1987). Monte carlo estimation under different distributions using the same simulation. *Technometrics*, 29(2):153–160.

Breeding, R. J., Helton, J. C., Gorhan, E. D., and Harper, F. T. (1992a). Summary description of the methods used in the probabilistic risk assessments for NUREG–1150. *Nuclear Engineering and Design*, 135:1–27.

Breeding, R. J., Helton, J. C., Murfin, W. B., and Smith, L. N. (1990). Evaluation of severe accident risks: Surry Unit 1. Technical Report NUREG/CR-4551, volume 3, Rev. 1–Pt. 1, U.S. Nuclear Regulatory Commission and Sandia National Laboratories.

Breeding, R. J., Helton, J. C., Murfin, W. B., Smith, L. N., Johnson, J. D., and Shiver, W. W. (1992b). The NUREG–1150 probabilistic risk assessment for the Surry nuclear power station. *Nuclear Engineering and Design*, 135:29–59.

Brown, T. D., Payne, A. C., Miller, L. A., Johnson, J. D., Chanin, D. I., Shiver, W. W., Higgins, S. J., and Sype, T. T. (1992). Integrated risk assessment for the LaSalle Unit 2 nuclear power plant. Technical Report NUREG/CR-5305, volume 1, U.S. Nuclear Regulatory Commission and Sandia National Laboratories.

Ericson, D. M., Wheeler, T. A., Sype, T. T., Drouin, M. T., Cramond, W. R., Camp, A. L., Maloney, K. J., and Harper, F. T. (1990). Analysis of core damage frequency: Internal events methodology. Technical Report NUREG/CR-4550, volume 2, U.S. Nuclear Regulatory Commission and Sandia National Laboratories.

Gorham, E. D., Breeding, R. J., Helton, J. C., Brown, T. D., Murfin, W. B., Harper, F. T., and Hora, S. C. (1993). Evaluation of severe accident risks: Methodology for the containment, source term, consequence, and risk integration analysis. Technical Report NUREG/CR-4551, volume 1, Rev. 1, U.S. Nuclear Regulatory Commission and Sandia National Laboratories.

Helton, J. C. and Breeding, R. J. (1993). Calculation of reactor accident safety goals. *Reliability Engineering and System Safety*, 39:129–158.

Iman, R. L. and Hora, S. C. (1990). A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis*, 10(3):401–406.

Jow, H.-N., Sprung, J. L., Rollstin, J. A., Ritchie, L. T., and Chanin, D. I. (1990). MELCOR accident consequence code system (MACCS). Technical Report NUREG/CR-4691, volume 2, U.S. Nuclear Regulatory Commission and Sandia National Laboratories, Albuquerque, NM.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics*, volume 2, chapter 26. MacMillan Publishing Co., New York, fourth edition.

McKay, M. D. (1995). Evaluating prediction uncertainty. Technical Report NUREG/CR–6311, U.S. Nuclear Regulatory Commission and Los Alamos National Laboratory.

McKay, M. D., Conover, W. J., and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

NUREG–1150 (1990). Severe accident risks: An assessment for five U.S. nuclear power plants. Technical Report NUREG–1150, U.S. Nuclear Regulatory Commission.

Parzen, E. (1962). *Stochastic Processes*, page 55. Holden Day, San Francisco.